Bagging-Based Active Learning Model for Named Entity Recognition with Distant Supervision

Sunghee Lee, Yeongkil Song, Maengsik Choi, Harksoo Kim
Program of Computer and Communications Engineering
College of IT, Kangwon National University
Chuncheon-si, Gangwon-do, Korea
{nlpfclee, nlpyksong, nlpmschoi, nlpdrkim}@kangwon.ac.kr

***Abstract*—Named entity recognition (NER) is a preliminary step to performing information extraction and question answering. Most previous studies on NER have been based on supervised machine learning methods that need a large amount of human-annotated training corpus. In this paper, we propose a semi-supervised NER model to minimize the time-consuming and labor-intensive task for constructing the training corpus. The proposed model generates weakly labeled training corpus using a distant supervision method. Then, it improves NER accuracy by refining the weakly labeled training corpus using a bagging-based active learning method. In the experiments, the proposed model outperformed the previous semi-supervised model. It showed F1-measure of 0.764 after 15 times of bagging-based active learning.**

Keywords—named entity recognition; distant supervision; bagging; active learning

I. INTRODUCTION

Named entities (NEs) are informative elements such as person's names, location's names, and organization's names in texts. Named entity recognition (NER) is a subtask of information extraction that finds NEs in texts and classifies the NEs into predefined classes such as PERSON, LOCATION, and ORGANIZATION, and locations, as shown in Fig. 1.

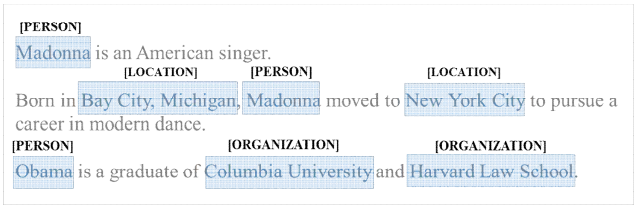

Fig.1. Example of named entity recognition

Previous NER systems are divided into two types: systems based on symbolic rules (rule-based systems) and systems based on machine learning (ML-based systems). Rule-based systems [1] use regular-expression-like patterns and NE dictionaries. If an NE dictionary is sufficiently large and patterns are generated by referring to a large corpus, the performances of rule-based systems may be satisfactory. However, it is well known that managing a lot of rules is very difficult and the cost for the initial implementation is high. ML-based systems [2-7] mainly use supervised learning models that collect statistical information from a large annotated corpus and determine NE classes based on this information. Recent ML-based systems have been focused on improving the accuracy of NER based on well-known supervised learning models such as DT (Decision Tree) [3], MEM (Maximum Entropy Model) [4], CRFs (Conditional Random Field) [5,6], and structural SVM (Support Vector Machine) [7]. The ML-based systems perform well, but the performances depend on the size of NE tagged training data. As well-known, it is a time-consuming and labor-intensive task to construct a large amount of NE tagged corpus. To reduce this problem, we propose a semi-supervised NER model using active learning [8] based on bagging (bootstrap aggregating) [9] with distant supervision [10]. The proposed model does not need a large amount of NE tagged training data and just needs an NE dictionary that contains NEs and their classes. By using a distant supervision method based on the NE dictionary, the proposed model automatically annotates raw corpus (i.e., a set of sentences that are not annotated with any additional tags) with NE classes. Then, it refines noises (i.e., words annotated with incorrect NE classes) by using bagging-based active learning and improves the accuracy of NER.

II. NAMED ENTITY RECOGNITION USING BAGGING-BASED ACTIVE LEARNING

A. Sytem Architecture

The proposed model consists of two parts; a distant supervision part that generates weakly labeled training corpus using a NE dictionary, and a bagging-based active learning part that improves NER accuracy by refining the weakly labeled training corpus, as shown in Fig. 2. In the distant supervision part, the proposed model automatically annotates a large amount of raw corpus with NE classes by simply matching word sequences against entries in a NE dictionary. In the bagging-based active learning part, the proposed model selects noise sentences from the weakly labeled training corpus based on disagreement scores between bagging models trained by the weakly labeled training corpus. Then, the noise sentences are manually revised according to an active learning method. This refinement process is repeated until some terminal conditions are satisfied.

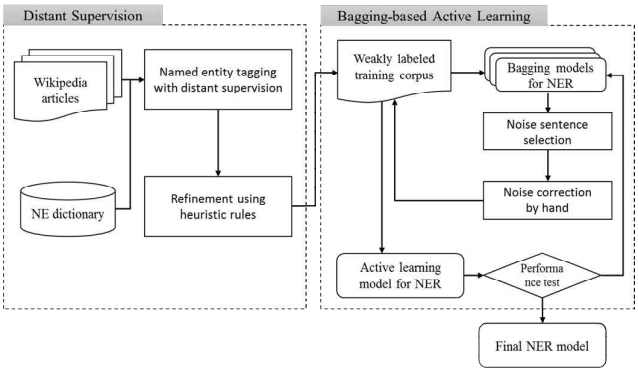

Fig.2. Overall architecture of the proposed model

B. Constructing Weakly Labeled Training Corpus Using Distant Supervision

For distant supervision, we use an NE dictionary that is semi-automatically constructed from Wikipedia (http://ko.wikipedia.org) [11]. The NE dictionary includes some noise entries (*i.e.*, the construction accuracy of the NE dictionary is the average micro F1-measure of 0.955) that are annotated with incorrect NE classes. Each entry in the NE dictionary is assigned into one NE class. In other words, although 'White House' can be location's name and organization's name according to its context, it is stored as either location's name or organization's name in the NE dictionary. Then, we collect articles from Wikipedia. After these preparations for distant supervision, the proposed model constructs weakly labeled training corpus by simply matching each sentence in the articles against entries in the NE dictionary. Then, the proposed model removes incorrect labels by using some heuristic rules, as follows.

(1) Remove the labels of words with declined or conjugated endings because endings of NEs are generally nouns.

(2) Remove labels of high frequent words in the weakly labeled training corpus because NEs are not common words as well-known in Zipf's law [12]. We remove the labels of words that occur in 10 times or more in the weakly labeled training corpus.

The heuristic rules can be changed according to target languages. Fig. 3 shows some Korean sentences that are weakly labeled by distant supervision with heuristic rules.

Fig.3. Example of weakly labeled sentences

In Fig. 3, we use Romanized Korean characters called *Hangeul*.

C. Refining Weakly Labeled Training Corpus Using Bagging-Based Active Learning

After constructing the weakly labeled training corpus, the proposed model refines noises using bagging-based active learning and improves accuracy of NER, as shown in Fig. 4, the bagging-based active learning algorithm.

(1) Generate *n* bagging corpus from training corpus by sampling with replacement.
(2) Train *n* NER models using *n* bagging corpus, respectively.
(3) Check disagreement scores between outputs of *n* NER models by using the whole training corpus as test data.
(4) Select *m* sentences with high disagreement scores.
(5) Revise incorrect labels in *m* sentences by hand.
(6) Update the training corpus with the revised sentences.
(7) Train an NER model using the updated training corpus.
(8) Check accuracy of the NER model by using gold-labeled test corpus.
(9) If accuracy improvement is converged, terminate the learning process. Otherwise, go to step (1).

Fig. 4. Algorithm of bagging-based active learning

In Fig. 4, we set *n* to a tenth of the whole training corpus. The disagreement scores are calculated by the numbers of NER models that return different NER results. In other words, if a sentence has two named entities in which the first NE is annotated with exclusively different labels by three NER models and the second NE is annotated with exclusively different labels by five NER models, a disagreement score of the sentence is five.

The NER models in Fig. 4 use CRFs [13] as a machine learning model and annotate input sentences according to the typical BIO (Begin-Inner-Outer) tagging scheme. For example, "Obama lives in White House" is labeled as "Obama/B_PER lives in White/B_LOC House/I". 'B_PER' means the beginning of person's name, and 'B_LOC' means the beginning of location's name. 'I' means the inner of a named entity. Table I shows input features of the NER models. As shown in Table I, the input features are designed for Korean sentences, but we believe that language change will not be a difficult task because the features are based on shallow NLP (natural language processing) knowledge.

TABLE I. LIST OF INPUT FEATURES FOR NER

Feature Name	Explanation
LEX	The current *eojeol* (Korean spacing unit)
FW_2_Lex, BW_2_Lex	First two *eomjeol*'s (Korean syllable) and last two *eomjeol*'s in the preceding, current, and next *eojeol*'s
FW_2_Tags, BW_2_Tags	NE categories matching against FW_2_Lex and BW_2_Lex in the preceding, current, and next *eojeol*'s
FW_3_Lex, BW_3_Lex	First three *eomjeol*'s and last three *eomjeol*'s in the preceding, current, and next *eojeol*'s

FW_3_Tags, BW_3_Tags	NE categories matching against FW_3_Lex and BW_3_Lex in the preceding, current, and next *eojeol*'s
BIEF (BE, BF, IE, IF)	BE: a tag meaning "the current *eojeol* is exactly matched against an entry in an NE dictionary" BF: a tag meaning "the current *eojeol* is partially matched against first few *eomjeol*'s in an entry in an NE dictionary" IE: a tag meaning "the current *eojeol* is included in an entry in an NE dictionary" IF: a tag meaning "the current *eojeol* is partially matched against last few *eomjeol*'s in an entry in an NE dictionary"
POS_Bigram	POS (part-of-speech) *bi*-grams of the preceding, current, and next *eojeol*'s
LEX-POS_Unigram	'Morpheme/POS' *uni*-grams of the preceding, current, and next *eojeol*'s

III. Evaluation

A. Data Sets and Experimental Settings

We collected 56,000 sentences including NEs from the Korean version of Wikipedia. Then, we divided the articles into training corpus (55,000 sentences) and testing corpus (1,000 sentences). The training corpus was used for distant supervision and bagging-based active learning. The testing corpus was manually annotated with gold labels (*i.e.*, correct NE classes). For experiments, we defined 11 NE classes: PERSON, LOCATION, ORGANIZATION, CELESTIAL_BODY, EVENT, FACILITY, GAME, LANGUAGE, LAW, PERSON_FICTION, and STUDY_FIELD. We set the threshold value of disagreement score to 10. Then, we performed 15 times of bagging-based active learning.

B. Experimental Results

The first experiment we conducted in this study was to evaluate the efficiency of the proposed model according to the number of iterations of bagging-based active learning, as shown in Table II. As shown in Table II, the proposed model gradually increased the F1-measure.

Table II. Performances at various iterations

# of iterations	Precision	Recall rate	F1-measure
1	0.868	0.650	0.744
5	0.868	0.675	0.759
10	0.860	0.683	0.761
15	0.862	0.687	0.764

In the second experiment, we compared the performances of the proposed model, as shown in Table III.

Table III. Performance comparison of NER models

Model	Precision	Recall rate	F1-measure
Kim [14]	0.726	0.628	0.674
DS	0.859	0.637	0.732
DS+BAL	0.862	0.687	0.764

In Table III, *Kim* [14] is a NER model using a typical distant supervision method without any refinement processes of weakly labeled corpus. *DS* is a NER model using the proposed distant supervision method with heuristic rules. *DS+BAL* is the proposed NER model using distant supervision and bagging-based active learning. As shown in Table III, the proposed models outperformed the previous semi-supervised NER model with distant supervision. In addition, we indirectly found that the heuristic rules and the bagging-based active learning contributed to remove noises in weakly labeled training corpus.

We found two major reasons that the proposed model failed to return correct NEs. First, incorrect entries in the NE dictionary caused wrong annotations in weakly labeling using distant supervision. Second, some NEs that were not in the NE dictionary were not participated in the training process. We think that these two problems can be overcome by refining the NE dictionary.

IV. Conclusion

We proposed a semi-supervised NER model based on distant supervision and bagging-based active learning. The proposed model effectively generates weakly labeled training corpus by using a distant supervision method with heuristic rules. Then, it improves the NER performances by using an ensemble method of bagging and active learning. In the experiments on NER of 11 NE classes, the proposed model showed better performances (precision of 0.862, recall rate of 0.687, F1-measure of 0.764) than the previous semi-supervised NER model. Based on the experimental results, we think that the proposed model may be a good solution to reduce the time-consuming tasks of training data construction because it does not require a large amount of NE tagged training corpus.

Acknowledgment

This work was supported by NCSOFT Corporation and ATC (Advanced Technology Center) Program "Development of Conversational Q&A Search Framework Based on Linked Data: Project No. 10048448". It was also supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2013R1A1A4A01005074).

References

[1] T. Noh and S. Lee, "Extraction and Classification of Proper Nouns by Rule-based Machine Learning," in Proceedings of the KIISE Korea Computer Congress 2000, Vol.27, No.2, 2000, pp. 170-172.

[2] Y. Hwang, H. Lee, E. Chung, B. Yun and S. Park, "Korean Named Entity Recognition Based on Supervised Learning Using Named Entity Construction Principles," in Proceedings of the HCLT, 2002, pp. 292-299.

[3] S. Sekine, R. Grishman and H. Shinnou, "A Decision Tree Method for Finding and Classifying Names in Japanese Texts," in Proceedings of 6th Workshop on Vary Large Corpora, 1998.

[4] A. Borthwick, J. Sterling, E. Agichtein and R. Grishman, "NYU: Description of the MENE Named Entity System as Used in MUC-7," in Proceedings of the Seventh Message Understanding Conference, 1997.

[5] W. W. Cohen and S. Sarawagi, "Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods," in Proceedings of the tenth ACM SIGKDD

international conference on Knowledge discovery and data mining, 2004, pp. 89-98.

[6] C. Lee, Y. Hwang, H. Oh, S. Lim, J. Heo, C. Lee, H. Kim, J. Wang and M. Jang, "Fine-Grained Named Entity Recognition using Conditional Random Fields for Question Answering," in Proceedings of the HCLT, 2006, pp. 268-272.

[7] C. Lee and M. Jang, "Named Entity Recognition with Structural SVMs and Pegasos Algorithm," Korean Journal of Cognitive Science, Vol. 21, No. 4, 2010, pp. 655-667.

[8] D.A. Cohn, Z. Ghahramani and M.I. Jordan, "Active learning with statistical models," Journal of artificial intelligence research, 1996.

[9] K. Ha, S. Cho, and D. MacLachlan, "Response models based on bagging neural networks," Journal of Interactive Marketing, Vol.19, No.1, 2005, pp. 17-30.

[10] M. Mintz, S. Bills, R. Snow and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Vol.2, 2009, pp. 1003-1011.

[11] Y. Song and H. Kim, "Semi-automatic Construction of a Named Entity dictionary Based on Active Learning," in Proceddings of the Computer Science and its Applications Lecture Notes in Electrical Engineering, Vol.330, 2015, pp. 65-70.

[12] G. K. Zipf, The Psychobiology of Language, Houghton-Mifflin, 1935.

[13] J. Lafferty, A. McCallum and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proceedings of the ICML, 2001, pp. 282-289.

[14] Y. Kim, Automatic training corpus generation method of Named Entity Recognition using Big data, M.S. Thesis, Sogang University, 2015.